University of Mumbai

Examination 2020 under cluster __(Lead College: _____)

Examinations Commencing from 7th January 2021 to 20th January 2021

Program: **Computer Engg**

Curriculum Scheme: Rev2016

Examination: BE Semester VII

Course Code: CSDLO7032

Course Name: Big Data Analytics

Time: 2 hour Max. Marks: 80

==========

==========

Q1.	**Choose the correct option for following questions. All the Questions are compulsory and carry equal marks**
1.	Which of the following combination is incorrect?
Option A:	Continuous – euclidean distance
Option B:	Continuous – correlation similarity
Option C:	Binary – manhattan distance
Option D:	None of the mentioned
2.	The goal of clustering a set of data is to
Option A:	predict the class of data
Option B:	divide them into groups of data that are near each other
Option C:	choose the best data from the set
Option D:	determine the nearest neighbors of each of the data
3.	The most important part of ___ is selecting the variables on which clustering is based.
Option A:	interpreting and profiling clusters
Option B:	selecting a clustering procedure
Option C:	assessing the validity of clustering
Option D:	formulating the clustering problem
4.	Find cosine distance between x and y X=[1,2,-1]. Y=[2,1,1]
Option A:	3/5
Option B:	1/2
Option C:	1/3
Option D:	4/5
5.	According to the original paper, which data-structure is used to store and maintain the points in CURE algorithm?
Option A:	hyperloglog
Option B:	k-d tree
Option C:	bloom filters
Option D:	hashmaps
6.	. Which of the following is a way to measure importance of web pages?
Option A:	PageRank

Option B:	Jaccard Distance
Option C:	DGIM algorithm
Option D:	Collaborative Filtering
7.	Which are the two types of hierarchical clustering?
Option A:	Fuzzy C-Means and Divise
Option B:	Divise and Agglomerative
Option C:	K-means and K-medoid
Option D:	Agglomerative and Fuzzy C-Means
8.	Which are the parameters used in graph?
Option A:	Centrality, Degree, Number of networks, Density
Option B:	Geodesic Distance, Degree, Density, Number of Vertices
Option C:	Centrality, Degree, Density, Geodesic Distance
Option D:	Direction, Degree, Density, Centrality
9.	In basic Page-Rank calculation Beta-(β) is also known as _____.
Option A:	Target PageRank
Option B:	Web structure ratio
Option C:	Teleportation Factor
Option D:	Page ratio
10.	Which of the following has a negative impact and will disturb the Page Rank system completely?
Option A:	Link spam
Option B:	Link spam farm
Option C:	Spam mass
Option D:	Trust ranking
11.	While measuring similarity in collaborative filtering which of the following distance calculations can be used?
Option A:	Jaccard and Manhattan Distance
Option B:	Jaccard and Cosine Distance
Option C:	Cosine and Hamming Distance
Option D:	Jaccard and Hamming distance
12.	In Collaborative filtering which of the following is the process of converting low ratings to negative value and high ratings to positive value?
Option A:	Normalizing Ratings
Option B:	Rounding the Ratings
Option C:	Measuring the Ratings
Option D:	Re-organizing the Ratings
13.	Which of the following is one of the major components of Flajolet martin algorithm?
Option A:	Length of the bit-string is $2^n<n$
Option B:	It deals with time stamps

Option C:	It consists of a collection of hash functions
Option D:	It is represented using $O(\log^2 N)$ bits
14.	Which of the following is not a component of Bloom filter?
Option A:	A set H of k hash functions each of which maps a key to one the n bits
Option B:	An array of n bits initialized to 1's
Option C:	A set S consisting of m number of key.
Option D:	All of the above
15.	Which of the following statements about the standard DGIM algorithm are false?
Option A:	DGIM operates on a time-based window.
Option B:	DGIM reduces memory consumption through a clever way of storing counts.
Option C:	In DGIM, the size of a bucket is always a power of two.
Option D:	The maximum number of buckets has to be chosen beforehand.
16.	Which of the following statements about the standard DGIM algorithm are false?
Option A:	DGIM operates on a time-based window.
Option B:	DGIM reduces memory consumption through a clever way of storing counts.
Option C:	In DGIM, the size of a bucket is always a power of two.
Option D:	The maximum number of buckets has to be chosen beforehand.
17.	A Bloom filter guarantees no _____ .
Option A:	False positives
Option B:	False negatives
Option C:	False positives and false negatives
Option D:	False positives or false negatives, depending on the Bloom filter type
18.	All types of processing such as sampling, cleaning, filtering, and querying on the input stream data is done in ____ .
Option A:	Input streams
Option B:	Stream processor
Option C:	Working storage
Option D:	Output streams
19.	Number of buckets required in DGIM algo is given by ____ .
Option A:	O(logN)

Option B:	O(log^2N)
Option C:	O(log2N)
Option D:	O(log3N)
20.	The FM algorithm can be used to
Option A:	Estimate the number of distinct elements
Option B:	Sample data with a time-sensitive window.
Option C:	Estimate the frequent elements.
Option D:	Determine whether an element has already occurred in previous stream data.

Q2. **(20 Marks)**	**Solve any Two Questions out of Three 10 marks each**
A	What are different types of distance measures?
B	Explain architectural patterns of NOSQL?.
C	Explain architecture of Hive?

Q3. **(20 Marks)**	
A	**Solve any Two 5 marks each**
i.	What are thr four V's of Bigdata?
ii.	Explain Combiners?
iii.	What is difference between traditional data and bigdata?
B	**Solve any One** **10 marks each**
i.	Explain hadoop ecosystem in detail?
ii.	Explain bloom filter analysys?